The Annals of Statistics
1998, Vol. 26, No. 3, 943–971

DIRECT ESTIMATION OF LOW-DIMENSIONAL COMPONENTS IN ADDITIVE MODELS[1]

BY JIANQING FAN,[2] WOLFGANG HÄRDLE AND ENNO MAMMEN

University of North Carolina, Humboldt-Universität zu Berlin and Ruprecht-Karls-Universität Heidelberg

Additive regression models have turned out to be a useful statistical tool in analyses of high-dimensional data sets. Recently, an estimator of additive components has been introduced by Linton and Nielsen which is based on marginal integration. The explicit definition of this estimator makes possible a fast computation and allows an asymptotic distribution theory. In this paper an asymptotic treatment of this estimate is offered for several models. A modification of this procedure is introduced. We consider weighted marginal integration for local linear fits and we show that this estimate has the following advantages.

(i) With an appropriate choice of the weight function, the additive components can be efficiently estimated: An additive component can be estimated with the same asymptotic bias and variance as if the other components were known.

(ii) Application of local linear fits reduces the design related bias.

1. Introduction. In this paper we consider the multivariate regression model

$$E(Y \mid X = x) = \mu + f_1(x_1) + f_{23}(x_2, x_3), \tag{1.1}$$

where Y is a real-valued dependent variable, $X = (X_1, X_2, X_3)$ is a vector of explanatory variables and μ is a constant. The variables X_1 and X_2 are continuous with values in $\mathbb{R}^p$ or $\mathbb{R}^q$, respectively, and X_3 is discrete and takes values in $\mathbb{R}^r$. For identifiability, we assume $Ef_1(X_1) = Ef_{23}(X_2, X_3) = 0$. The novelty of this paper is to directly estimate $f_1(x)$ at the usual nonparametric rate with good sampling properties. Our model includes the additive nonparametric regression model:

$$E(Y \mid U = u) = \mu + g_1(u_1) + \cdots + g_p(u_p), \tag{1.2}$$

where now $U = (U_1, \ldots, U_p)$ is a vector of explanatory variables. A discussion of this model can be found in Buja, Hastie and Tibshirani (1989) and Hastie and Tibshirani (1990). Model (1.2) is easy to interpret and is much more flexible than a linear model. Furthermore, the additive components g_j can be estimated with the one-dimensional nonparametric rate [Stone (1985, 1986)].

Received December 1995; revised August 1997.

[1] This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse," Humboldt-Universität zu Berlin and NATO Collaborative Research Grant 931 363.

[2] Supported in part by NSF Grant DMS-95-04414 and NSA grant 96-1-0015.

The main conclusion of this paper is somewhat surprising: The component g_j can be estimated with the same asymptotic bias and variance as the one-dimensional smoother, as if the other components were known. This kind of adaptivity result appears to be new in the literature. It provides foundational insights into additive modeling: Unknown components in the additive model, although increasing the effective number of parameters, do not add any extra difficulty of estimation, at least asymptotically.

In most papers, for the calculation of the additive components, algorithms have been proposed which are based on iterative procedures using backfitting. Recently, asymptotic properties of backfitting estimates have been analyzed in Opsomer and Ruppert (1997), Opsomer (1997) and Linton, Mammen and Nielsen (1997). Because of the implicit definition of these estimates, their behavior is difficult to understand. For this reason, in Linton and Nielsen (1995), Tjøstheim and Auestad (1994) and Chen, Härdle, Linton and Severance-Lossin (1996) a direct method has been proposed that is based on "marginal integration." This procedure is based on the fact that, up to a constant, $g_j(u_j)$ is equal to

$$EW(U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_s)m(U_1, \ldots, U_{j-1}, u_j, U_{j+1}, \ldots, U_s),$$

where $m(u) = E(Y \mid U = u)$. Here W is a weight function with

$$EW(U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_s) = 1.$$

The estimate of g_j is achieved by (weighted) marginal integration of an estimate of m. In particular, this method does not use iterations. Fast computation can be implemented. Furthermore, the explicit definition allows a detailed asymptotic analysis.

The present paper extends this idea in two directions:

(i) It introduces a weighting scheme W, which leads to efficient estimation [for another proposal for efficient estimation based on marginal integration, see Linton (1997)].

(ii) It allows a more flexible model, which can be incorporated with discrete data.

Our asymptotic analysis can be extended to the case that model (1.1) does not hold (see Remark 3). Then in the case of the additive model (1.2) the marginal integration estimate gives a consistent estimate of

$$\bar{g}_j(u_j) = EW(U_1, \ldots, U_{j-1}, U_{j+1}, \ldots, U_s)m(U_1, \ldots, U_{j-1}, u_j, U_{j+1}, \ldots, U_s).$$

This can be interpreted as an average effect of the jth component and is the best additive approximation under some specific L_2-norm [see Fan (1997)]. The backfitting estimate behaves quite differently. Under appropriate conditions it is a consistent estimate of g_j^* where $\mu + g_1^*(u_1) + \cdots + g_p^*(u_p)$ is the orthogonal projection in the Hilbert Space $L_2(p)$ onto the subspace of additive functions. Here p is the joint density of $(U_1, \ldots, U_p)$ (design density). For identifiability, g_j^* is normed s.t. $Eg_j^*(U_j) = 0$. This statement follows from the results of Linton, Mammen and Nielsen (1997). So, if model (1.2) is only

approximately true, we conjecture that backfitting will lead to a more accurate estimate of the full-dimensional regression function m. This would be preferable if one is interested in prediction. Furthermore, the application of marginal integration requires consistency of a full-dimensional smoother. This puts restrictions on the dimension that may not be shared by the backfitting estimate; see Linton, Mammen and Nielsen (1997). On the other hand, (in the case of model misspecification), the average effect $\bar{g}_j$ is always easy to interpret and it may be argued that marginal integration is preferable as a data analytic tool.

Our model includes additive partial linear models. With $X = (U_1, \ldots, U_p, X_3)$, $x = (u_1, \ldots, u_p, x_3)$ we write

$$E(Y \mid X = x) = \mu + g_1(u_1) + \cdots + g_p(u_p) + x_3^T \beta. \tag{1.3}$$

In this case, each nonparametric additive component can be estimated with optimal rate by our direct estimate $\hat{g}_j$, $j = 1, \ldots, p$. Furthermore, we will show that a least-squares estimate

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta} \end{pmatrix} = (Z^T Z)^{-1} Z^T \left(Y - \hat{g}_1 - \cdots - \hat{g}_p\right)$$

possesses root-n consistency. Here, for n observations $Y_1, \ldots, Y_n$ and design vectors $X^i = (U_{1i}, \ldots, U_{pi}, X_{3i})$, $i = 1, \ldots, n$, the vectors Y and $\hat{g}_j$ have elements Y_i and $\hat{g}_j(U_{ij})$, respectively, $i = 1, \ldots, n$; $j = 1, \ldots, p$. The design matrix Z has rows $(1, X_{3i}^T)$.

Another application of our model consists of partial interaction models

$$E(Y \mid U = u) = \mu + g_{12}(u_1, u_2) + g_3(u_3) + \cdots + g_s(u_s). \tag{1.4}$$

Our method directly applies interactions such as g_{12} by treating the rest of the variables as X_2-vectors and/or X_3-vectors [see (1.1)].

This paper is organized as follows. In Section 2, we introduce our estimation procedure. Section 3 presents asymptotic results. A further discussion of additive models (1.2), additive partially linear models (1.3) and partial interaction models (1.4) can be found in Section 4. In Section 5 our methodology is applied to a data set on female labor supply in East Germany. Furthermore, there a small simulation study can be found. Assumptions and proofs are postponed to Section 6.

2. Estimation procedure. Let $m(x_1, x_2, x_3) = E(Y \mid X_1 = x_1, X_2 = x_2, X_3 = x_3)$ be the regression function and let $W\colon \mathbb{R}^{q+r} \to \mathbb{R}$ be a known function with $EW(X_2, X_3) = 1$. Observe that under (1.1)

$$\begin{aligned} Em(x_1, X_2, X_3)W(X_2, X_3) &= \mu + f_1(x_1) + Ef_{23}(X_2, X_3)W(X_2, X_3) \\ &= \mu_1 + f_1(x_1) \\ &\equiv f_1^*(x_1), \end{aligned} \tag{2.1}$$

where

$$\mu_1 = \mu + Ef_{23}(X_2, X_3)W(X_2, X_3). \tag{2.2}$$

Thus, f_1 can be directly estimated within a constant factor. This can be done by averaging out a nonparametric estimator of m with respect to other variables X_2, X_3. Since, in practice, $f_1(\cdot)$ will be normalized to have sample mean 0, the constant fact μ_1 is irrelevant to the final estimated curve. This kind of integration idea was studied in the additive model (1.1) by Tjøstheim and Auestad (1994), Linton and Nielsen (1995) and Chen, Härdle, Linton and Severance-Lossin (1996).

To utilize (2.1), we consider the local linear approximation near a fixed point x_1:

$$f_1(v_1) \approx a(x_1) + b^T(x_1)(v_1 - x_1),$$

where v_1 lies in a neighborhood of x_1. Further, the local constant approximation for f_{23} at a fixed point x_2 and x_3 is employed:

$$f_2(v_2, x_3) \approx c(x_2, x_3) \quad \text{for } v_2 \approx x_2.$$

Thus, in a neighborhood of (x_1, x_2) and for the given value of x_3, we can approximate the regression function as

$$\begin{aligned} m(v_1, v_2, x_3) &\approx \mu + a(x_1) + b^T(x_1)(v_1 - x_1) + c(x_2, x_3) \\ &\equiv \alpha + \beta^T(v_1 - x_1). \end{aligned} \tag{2.3}$$

Note that $f_{23}(\cdot; x_3)$ is locally approximated by a constant. This is because:

(i) the function $c(x_2, x_3)$ will be averaged out by an integration via (2.1);
(ii) the higher-order approximation will increase the number of local parameters and hence is harder to implement in higher dimensions.

In principle, we can approximate $f_1(\cdot)$ to a higher order. We opt not to do this for simplicity. Furthermore, the higher-order approximation rarely takes effect for the finite amount of data—the size of the local neighborhood plays a more crucial role [see, e.g., Fan and Gijbels (1996)].

Consider now that we have an i.i.d. data set $(Y_i, X_{1i}, X_{2i}, X_{3i})$, $i = 1, \ldots, n$, for model (1.1). The local model (2.3) leads to the following regression problem: Minimize

$$\sum_{i=1}^{n} \left(Y_i - \alpha - \beta^T(X_{1i} - x_1)\right)^2 K_{h_1}(X_{1i} - x_1) L_{h_2}(X_{2i} - x_2) I\{X_{3i} = x_3\}. \tag{2.4}$$

Here K and L are kernel functions and for bandwidths h_1 and h_2 we put

$$K_{h_1}(t) = \frac{1}{h_1^p} K\left(\frac{t}{h_1}\right) \quad \text{and} \quad L_{h_2}(t) = \frac{1}{h_2^q} L\left(\frac{t}{h_2}\right).$$

Note that the factor $K_{h_1} L_{h_2} I$ in (2.4) is just to confine our localization idea. Let $\hat{\alpha}(x)$ and $\hat{\beta}(x)$ be the solution to (2.4). Then, from (2.3) by setting $(v_1, v_2, x_3) = x$, we can easily see that $m(x) \approx \alpha$. Thus, our partial local

linear estimator is $\hat{m}(x) = \hat{\alpha}$. By (2.1), we propose the following estimator:

$$\hat{f}_1^*(x_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(x_1, X_{2i}, X_{3i}) W(X_{2i}, X_{3i}) \tag{2.5}$$

and

$$\hat{f}_1(x_1) = \hat{f}_1^*(x_1) - \bar{f}_1, \qquad \bar{f}_1 = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_1^*(X_{1i}). \tag{2.6}$$

Note that when the local constant fit is employed (i.e., $\beta = 0$) in (2.3), the resulting estimate $\hat{\alpha}$ is basically the multivariate kernel regression estimator.

Let X be the design matrix and let A be the diagonal weight matrix to the least-squares problem (2.4). Then

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X^T A X)^{-1} X^T A Y,$$

where $Y = (Y_1, \ldots, Y_n)^T$, and simple algebra shows that $\hat{m}(x) = \hat{\alpha}$ can be expressed as

$$\hat{m}(x) = \sum_{i=1}^{n} K_n(X_i - x) Y_i, \tag{2.7}$$

where, with $S_n(x) = (X^T A X)$ and $e_1^T = (1, 0, \ldots, 0)$,

$$K_n(t_1, t_2, t_3) = e_1^T S_n^{-1} \begin{pmatrix} 1 \\ t_1 \end{pmatrix} K_{h_1}(t_1) L_{h_2}(t_2) I\{t_3 = 0\}. \tag{2.8}$$

Note that it follows from least-squares theory that

$$\sum_{i=1}^{n} K_n(X_i - x) = 1 \quad \text{and} \quad \sum_{i=1}^{n} K_n(X_i - x)(X_{1i} - x_1) = 0. \tag{2.9}$$

3. Main results. Let us begin by introducing some notation. Let $p_1(x_1)$ and $p_{1,2}(x_1, x_2)$ be respectively the density of X_1 and (X_1, X_2) and let $p_{1,2|3}(x_1, x_2, | x_3)$, $p_{2|3}(x_2 | x_3)$ be respectively the conditional density of (X_1, X_2) given X_3 and of X_2 given X_3. Set $p_3(x_3) = P(X_3 = x_3)$. The conditional variance of $\varepsilon = Y - E(Y | X)$ is denoted by

$$\sigma^2(x) = E(\varepsilon^2 | X = x) = \operatorname{var}(Y | X = x),$$

where $X = (X_1, X_2, X_3)$. Let

$$\|K\|^2 = \int K^2 \quad \text{and} \quad \mu_2(K) = \int t t^T K(t)\, dt.$$

Then, under Condition A in Section 6, we have the following theorem that generalizes the main result in Linton and Nielsen (1995).

THEOREM 1. *Under Condition* A *for a point* $x_1 \in \mathbb{R}^p$, *if the bandwidths are chosen such that* $nh_1^p h_2^q/\log n \to \infty$, $h_1 \to 0$, $h_2 \to 0$ *in such a way that* $h_2^d/h_1^2 \to 0$, *then*

$$\sqrt{nh_1^p}\left\{\hat{f}_1^*(x_1) - f_1(x_1) - \mu_1 - b(x_1) + o\left(h_1^2\right)\right\} \to N(0, v(x_1)), \tag{3.1}$$

where

$$b(x_1) = \tfrac{1}{2}h_1^2 \operatorname{tr}\left(f_1''(x_1)\mu_2(K)\right) \tag{3.2}$$

and

$$\begin{aligned} v(x_1) &= \|K\|^2 p_1(x_1) \\ &\quad \times E\left\{\sigma^2(X_1, X_2, X_3)\frac{p_{2|3}^2(X_2 \mid X_3)W^2(X_2, X_3)}{p_{1,2|3}^2(X_1, X_2 \mid X_3)}\middle| X_1 = x_1\right\}. \end{aligned} \tag{3.3}$$

REMARK 1. Condition A(vi) is also not a necessary condition for Theorem 1. It is imposed to simplify the technical proof. In the proof we approximate the matrix S_n^{-1} by a deterministic sequence. If we used a higher-order stochastic expansion of S_n^{-1}, Condition A(vi) could be weakened. Note that if the local polynomial of order d is used to approximate the function f_2 instead of using the local constant fit with a higher-order kernel, then the result of Theorem 1 continues to hold without imposing Condition A(vi) and the derivative conditions on $p_{1,2|3}(x_1, x_2 \mid x_3)$. In other words, these conditions are not essential to our estimation problem.

REMARK 2. Under the additional assumptions that X_1 has compact support $\mathscr{X}$ and that Condition A holds uniformly for $x_1 \in \mathscr{X}$ [i.e., the infimum in A(iii) is uniformly bounded from below and the derivatives considered in A(iii) and A(iv) are uniformly bounded], it is easy to show that

$$\bar{f}_1 = \frac{1}{n}\sum_{i=1}^n \hat{f}_1^*(X_{1i}) = \bar{b} + o\left(h_1^2\right) + o_P\left(\frac{1}{\sqrt{nh_1^p}}\right),$$

where $\bar{b} = \frac{1}{2}h_1^2 E\,\operatorname{tr}[f_1''(X_1)\mu_2(K)]$. So it follows from Theorem 1 for $\hat{f}_1 = \hat{f}_1^* - \bar{f}_1$ that

$$\sqrt{nh_1^p}\left\{\hat{f}_1(x_1) - f_1(x_1) + \bar{b} - b(x_1) + o\left(h_1^2\right)\right\} \to N(0, v(x_1)).$$

Note that the "additional bias" term $\bar{b}$ can be dropped in the preceding expression if a different bandwidth (smaller than h_1) is used to construct $\bar{f}_1$. If one can only assume that Condition A holds uniformly over a subset $\mathscr{X}'$ of $\mathscr{X}$, then one could consider $\hat{f}_1 = \hat{f}_1^* - \bar{f}_1$ with $\bar{f}_1 = \sum_{i=1}^n \gamma(X_{1i})\hat{f}_1^*(X_{1i})/\sum_{i=1}^n \gamma(X_{1i})$, where γ is a weight function that vanishes outside of $\mathscr{X}'$. Then $\hat{f}_1$ is a consistent estimate of $f_1(x_1) - E\gamma(X_1)f_1(X_1)/E\gamma(X_1)$ and its asymptotic distribution can be easily seen from Theorem 1. Our following results have similar implications. For brevity we will not mention them.

REMARK 3. An analogous result can be proved for the case that model (1.1) does not hold, that is, that the regression function $m(x) = E[Y \mid X = x]$ is not of the form $\mu + f_1(x_1) + f_{23}(x_2, x_3)$. If Condition A(iv) is replaced by the assumption that $m(u_1, u_2, u_2)$ has bounded partial derivatives up to order 2 with respect to u_1 and up to order d with respect to u_2 for u_1 in a neighborhood of x_1 and for (u_2, u_3) in the support of the weight function W, one can show that (3.1) holds with

$$b_1(x_1) = \frac{1}{2} h_1^2 \mu_2(K) E\left\{ \operatorname{tr}\left[\frac{\partial^2}{(\partial x_1)^2} m(x_1, X_2, X_3) \right] W(X_2, X_3) \right\}$$

and with $\mu_1 + f_1(x_1)$ replaced by $E[m(x_1, X_2, X_3)W(X_2, X_3)]$. In this case $\hat{f}_1^*$ is a consistent estimate of a weighted average effect of the covariable X_1.

REMARK 4. Due to the local linear fitting, the resulting estimate $\hat{f}_1^*(x_1)$ is automatically adapted to the boundary of the design density of X_1. This can be seen from our proof. The theoretical formulation of boundary properties of a nonparametric estimator can be found in Gasser and Müller (1979) and its applications to the local polynomial fitting is given by Fan and Gijbels (1992, 1996), and Ruppert and Wand (1994).

We now consider the optimal weight function $W(\cdot)$. This is equivalent to minimizing

$$\min_W E\left\{ \sigma^2(X) \frac{p_{2\mid 3}^2(X_2 \mid X_3) W^2(X_2, X_3)}{p_{1,2\mid 3}^2(X_1, X_2 \mid X_3)} \mid X_1 = x_1 \right\} \tag{3.4}$$

subject to $EW(X_2, X_3) = 1$.

We first state a simple lemma.

LEMMA 1. *The minimization problem* $\min_W \int W^2(x) g^2(x)\, dx$ *subject to* $\int W(x)h(x) = 1$ *is obtained at*

$$W = \frac{h(x)}{g^2(x)} \Big/ \int \frac{h^2}{g^2}$$

and the minimum value is $\{\int h^2(x)/g^2(x)\, dx\}^{-1}$.

PROOF. Using the Language multiplier method, we have to minimize $\int W^2 g^2 - \theta W h$. This is equivalent to minimizing $W^2(x)g^2(x) - \theta W(x)h(x)$, yielding the solution

$$W(x) = \frac{\theta}{2} \frac{h(x)}{g^2(x)}.$$

The constraint $\int Wh = 1$ gives

$$W(x) = \frac{h(x)}{g^2(x)} \Big/ \int \frac{h^2}{g^2}.$$

This completes the proof. □

Applying Lemma 1 to problem (3.4), we obtain the optimal solution

$$W(X_2, X_3) = c^{-1} \frac{p_{2,3}(X_2, X_3)\, p^2_{1,2\mid 3}(x_1, X_2 \mid X_3)}{\sigma^2(x_1, X_2, X_3)\, p^2_{2\mid 3}(X_2 \mid X_3)\, p_{2,3\mid 1}(X_2, X_3 \mid x_1)} = c^{-1} \frac{p(x_1, X_2, X_3)\, p_1(x_1)}{\sigma^2(x_1, X_2, X_3)\, p_{2,3}(X_2, X_3)}, \tag{3.5}$$

where $p(x) = p_{1,2\mid 3}(x_1, x_2 \mid x_3) p_3(x_3)$ and $p_{2,3}(x) = p_{2\mid 3}(x_2 \mid x_3) p_3(x_3)$ are respectively the joint "density" of $X = (X_1, X_2, X_3)$ and (X_2, X_3) and where $c = p_1(x_1)^2 E\{\sigma^{-2}(X) \mid X_1 = x_1\}$. The minimal variance is

$$\min_W v(x_1) = \frac{\|K\|^2}{p_1(x_1)} \left[E\{\sigma^{-2}(X) \mid X_2 = x_1\} \right]^{-1}. \tag{3.6}$$

REMARK 5. The optimal weight function W depends on x_1. When it is used, the constant μ_1 [see (2.2)] depends on x_1. So in this case the estimate $\hat{f}_1^*(x_1)$ no longer estimates a function that is parallel to f_1. Nevertheless the estimate $\hat{f}_1$ is a consistent estimate of f_1. Note that for the calculation of $\hat{f}_1(x_1)$ the same weight function (depending on x_1) is used for $\hat{f}_1^*(X_{1i})$ in (2.6). Therefore the term $\mu_1 = \mu_1(x_1)$ cancels. See also Remark 2. Furthermore, as noted in Remark 2, the extra term of bias can be completely eliminated if a different bandwidth is applied to construct $\bar{f}$.

REMARK 6. Typically, the design densities $p(X)$, $p_1(X_1)$, $p_{2,3}(X_2, X_3)$ are not known. A theoretically satisfactory way out consists of dividing our sample into a relatively small first subsample and a relatively large second subsample. Then, under our smoothness assumptions, the design densities can be consistently estimated by the first subsample. The regression functions can be estimated in a second step using the other subsample. This shows that the optimal variance can be achieved, at least theoretically. The practically more relevant procedure, using the full data set for the estimation of the design densities and of the regression function, is not covered by our theory.

REMARK 7. When $f_{23}(x_2, x_3)$ is known and $\sigma^2(x) \equiv \sigma^2$, one can directly smooth $Y - f_{23}(X_2, X_3)$ on X_1 to obtain an estimate of $f_1(x_1)$ and this estimate is optimal in an asymptotic minimax sense [cf. Fan (1993)]. The variance of this estimate is $\sigma^2\|K\|^2/p_1(x_1)$, which is the same as (3.6). In other words, our direct estimator (2.6) shares the same optimality as this ideal estimator and has the same ability of estimating the additive component even if f_{23} is unknown.

REMARK 8. In the case that X_1 is independent of (X_2, X_3) and $\sigma^2(x) \equiv \sigma^2$, one can directly smooth Y on X_1 to obtain an estimate of $f_1(x_1)$ [cf. Härdle and Tsybakov (1995)]. This estimator has the asymptotic variance

$$\frac{\|K\|^2}{p_1(x_1)}\left[\sigma^2 + \operatorname{var}\{f_{23}(X_2, X_3)\}\right],$$

which is larger than our direct estimator (2.6) with the optimal weight (3.5).

To summarize, we have

THEOREM 2. *Under the assumptions of Theorem* 1, *if the ideal weight* (3.5) *is used*, *we have*

$$\sqrt{nh_1^p}\left\{\hat{f}_1^*(x_1) - f_1(x_1) - \mu_1 - b(x_1) + o\left(h_1^2\right)\right\} \to N\left(0, \frac{\|K\|^2}{p_1(x_1)}\left[E\left\{\sigma^{-2}(X) \mid X_1 = x_1\right\}\right]^{-1}\right),$$

where $b(x_1)$ *was defined in* (3.2).

4. Applications to special models.

4.1. *Additive model*. We now assume the following additive model:

$$Y = \mu + g_1(U_1) + \cdots + g_p(U_p) + \varepsilon, \tag{4.1}$$

where $g_1(\cdot), \ldots, g_p(\cdot)$ are univariate functions satisfying the identifiability condition

$$E_{g_1}(U_1) = 0, \ldots, E_{g_p}(U_p) = 0$$

and $U_1, \ldots, U_p$ are continuous variables having a joint density p. Now, for each variable U_α, we can form directly $\hat{g}_\alpha^*$ as in (2.6), using now $h_1 = h_{1\alpha}$ and $h_2 = h_{2\alpha}$.

THEOREM 3. *If the conditions of Theorem* 1 *hold for each component* α, *then we have the following joint asymptotic normality*:

$$\begin{pmatrix} \sqrt{nh_{11}}\left\{\hat{g}_1^*(u_1) - g_1(u_1) - \mu_{11} - \frac{1}{2}h_{11}^2\mu_2(K)g_1''(u_1) + o\left(h_{11}^2\right)\right\} \\ \vdots \\ \sqrt{nh_{1p}}\left\{\hat{g}_p^*(u_p) - g_p(u_p) - \mu_{1p} - \frac{1}{2}h_{1p}^2\mu_2(K)g_p''(u_1) + o\left(h_{1p}^2\right)\right\} \end{pmatrix} \to_d N(0, \Sigma), \tag{4.2}$$

where $\mu_{1\alpha}$ *is analogous to that defined in* (2.1) *and*

$$\Sigma = \|K\|^2 \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$$

and

$$\sigma_\alpha^2(u) = E\left\{\frac{\sigma^2(U)p_\alpha(u_\alpha)p_{-\alpha}^2(U_{-\alpha})W_\alpha^2(U_{-\alpha})}{p^2(U)}\middle| U_\alpha = u_\alpha\right\},$$

with $U_{-\alpha} = (U_1, \ldots, U_{\alpha-1}, U_{a+1}, \ldots, U_p)$ *and* $p_{-\alpha}$ *is its joint density*.

REMARK 9. When $W \equiv 1$, the variance matrix Σ is the same as that obtained by Chen, Härdle, Linton and Severance-Lossin (1996). However, since we employ the local linear fit, our bias has a nicer expression. Put another way, the local linear fit (2.6) uses one extra local parameter without increasing the variance. See Fan and Gijbels (1996) for further discussion on the advantages of using local polynomial fits.

REMARK 10. Under the standard assumption that all components are only two times continuously differentiable (i.e., $d = 2$) and smoothing of optimal order is done for α (i.e., $h_{1\alpha}$ is of order $n^{-1/5}$), then the conditions $nh_{1\alpha}h_{2\alpha}^{p-1}/\log n \to \infty$, $h_{2\alpha}/h_{1\alpha} \to 0$ imply $p \le 4$. [Furthermore, Condition A(vi) implies $p \le 2$. However, this condition can be weakened; see Remark 1.] So for $p \ge 5$ two times differentiable component rates of order $n^{-2/5}$ cannot be achieved by the marginal integration estimate. However, with a modification given by Hengartner (1996), the marginal integration estimate can still achieve the optimal rate of convergence.

If the ideal weight scheme (3.5) is applied to each additive component, the weight function should be

$$W_\alpha = \frac{p(U)\,p_\alpha(U_\alpha)}{\sigma^2(U)\,p_{-\alpha}(U_{-\alpha})} \Bigg/ \int \frac{p(U)\,p_\alpha(U_\alpha)}{\sigma^2(U)}\,dU_{-\alpha} \tag{4.3}$$

and the ideal variance is $\|K\|^2\sigma^2/p_\alpha(U_\alpha)$ if $\sigma^2(U) \equiv \sigma^2$.

4.2. *Additive partially linear model.* Consider the additive partially linear model (1.3), which possesses the flexibility to model a part of covariates (in particular, discrete variables) linearly. In this model, one can form the estimate of $g_\alpha(\cdot)$ via $\hat{g}_\alpha^*(\cdot)$ as in Section 4.1 [by treating the additional discrete variable X_3 as in Section 3]. Let

$$\theta = \sum_{\alpha=1}^{p} \mu_{1\alpha}. \tag{4.4}$$

Then $\hat{g}_1^* + \cdots + \hat{g}_p^*$ overestimates $g_1 + \cdots + g_p$ by an amount of θ. Since model (1.3) involves an intercept term, this will only affect the estimate of μ, not the slope β. Since the grand mean $\mu = EY - EX_2^T\beta$ can be estimated as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} X_{3i}^T\hat{\beta}, \tag{4.5}$$

the actual value of θ is not a concern to us.

The quality of the estimator $\hat{g}_\alpha^*$ is not high at the region where the data are sparse. To eliminate such deficiencies used in the parametric estimation, we use the ith data point if $X^i = (U_{1,i}, \ldots, U_{p,i}, X_{3,i}) \in A$, where A is a

prescribed set (usually a rectangle) in $\mathbb{R}^{p+r}$. Now consider the following least-squares problem:

$$\min_{\mu,\beta} \sum_{i=1}^{n} \left\{ Y_i - \hat{g}_1^*(U_{1i}) - \cdots - \hat{g}_p^*(U_{pi}) - \mu - X_{3i}^T \beta \right\}^2 I\{X^i \in A\}. \tag{4.6}$$

Let $Z_i = \begin{pmatrix} 1 \\ X_{3i} \end{pmatrix}$ and $\tilde{Z} = (Z_1, \ldots, Z_n)^T$ be the design matrix.

Put $\Delta = \operatorname{diag}\{I\{X^1 \in A\}, \ldots, I\{X^n \in A\}\}$ and $\hat{\beta}^* = \begin{pmatrix} \hat{\mu} \\ \hat{\beta} \end{pmatrix}$. Then

$$\hat{\beta}^* = \left(\tilde{Z}^T \Delta \tilde{Z}\right)^{-1} \tilde{Z}^T \Delta \left(Y - \hat{g}_1^* - \cdots - \hat{g}_p^*\right),$$

where $Y = (Y_1, \ldots, Y_n)^T$ and $\hat{g}_\alpha^* = (g_\alpha^*(U_{\alpha 1}), \ldots, g_\alpha^*(U_{\alpha n}))^T$. To state the asymptotic normality of $\hat{\beta}^*$, we use the notation introduced in Section 4.1. Additionally, we need the following notation.

Let $p_\alpha(\cdot)$ be the marginal density of U_α and let $p_{-\alpha,3}(\cdot)$ be the marginal density of $(U_{-\alpha}, X_3)$, $\alpha = 1, \ldots, p$,

$$Z_A = Z I(X \in A) - \sum_{\alpha=1}^{p} \frac{W_\alpha(U_{-\alpha}, X_3)\, p_{-\alpha,3}(U_{-\alpha}, X_3)}{p(X)} p_\alpha(U_\alpha)\, E\{Z I(X \in A) \mid U_\alpha\}.$$

Put $\beta^* = \begin{pmatrix} \mu + \theta \\ \beta \end{pmatrix}$. For simplicity of discussion, we assume that $W_\alpha(\cdot)$ is independent of u_α. (Otherwise, the root-n of $\hat{\beta}^*$ holds, but the covariance is more complicated. Set

$$\begin{aligned} V_\alpha = {} & \{W_\alpha(U_{-\alpha}, X_3) - 1\} E\left[g_\alpha(U_\alpha) I\{X \in A\} Z\right] \\ & + \Big\{ \left(g_{-\alpha}(U_{-\alpha}) + X_3^T \beta\right) W_\alpha(U_{-\alpha}, X_3) \\ & \quad - E\left[\left(g_{-\alpha}(U_{-\alpha}) + X_3^T \beta\right) W_\alpha(U_{-\alpha}, X_3)\right] \Big\} E\left[I\{X \in A\} Z\right], \end{aligned}$$

where

$$g_{-\alpha}(U_{-\alpha}) = g_1(U_1) + \cdots + g_{\alpha-1}(U_{\alpha-1}) + g_{\alpha+1}(U_{\alpha+1}) + \cdots + g_p(U_p).$$

THEOREM 4. *Under the assumptions of Theorem* 3, *if* $\|X_3\|$ *has a bounded fourth moment*, $nh_{1\alpha}^2 h_{2\alpha}^{2(p-1)}/(\log n)^2 \to \infty$ *and* $h_{1\alpha} = o(n^{-1/4})$, *we have*

$$\sqrt{n}\left(\hat{\beta}^* - \beta^*\right) \to N\left(0, B_1^{-1} B_2 B_1^{-1}\right),$$

where

$$B_1 = E I(X \in A) Z Z^T$$

and

$$B_2 = E\sigma^2(X) Z_A Z_A^T + \operatorname{var}\left(\sum_{\alpha=1}^{p} V_\alpha \right).$$

When X_3 contains quite a few binary variables, the estimator (2.6) can be hard to use, since few data points are available in (2.4). For the additive

partially linear model (1.3), special care is needed. In the local step, we can replace (2.4) by

$$(4.7)\quad \sum_{i=1}^{n} \left(Y_i - a - b(U_{1i} - u_1) - X_{3i}^T \beta\right)^2 K_{h_1}(U_{1i} - u_1) L_{h_2}(X_{2i} - x_2),$$

where $X_{2i} = (U_{2i}, \ldots, U_{pi})^T$ and $x_2 = (u_2, \ldots, u_p)^T$. Note that (4.7) is obtained via the local regression model in a neighborhood of (u_1, x_2). This kind of idea appears already in Carroll, Fan, Gijbels and Wand (1997). We denote $g(u) = g_1(u_1) + \cdots + g_p(u_p)$. Let $\hat{a}$, $\hat{b}$ and $\hat{\beta}$ minimize (4.7). Then

$$\hat{g}^*(u_1, x_2) = \hat{a}$$

is a nonparametric estimator of g. Let $W(x_2)$ be a function such that $EW(X_2) = 1$ and

$$g_1^+(u_1) = \mu + g_1(u_1) + EW(X_2) f_2(X_2) = g_1(u_1) + \mu_1^*,$$

where $f_2 = g_2 + \cdots + g_p$. Then

$$(4.8)\quad \hat{g}_1^+(u_1) = n^{-1} \sum_{i=1}^{n} \hat{g}^*(u_1, X_{2i}) W(X_{2i})$$

is an estimator of $g_1^+(u_1)$, with the following asymptotic properties.

THEOREM 5. *Suppose that Condition* B *holds for* $\alpha = 1$. *Then, if* $nh_1 h_2^{p-1}/\log n \to \infty$ *and* $h_1 \to 0$ *and* $h_2^d/h_1^2 \to 0$,

$$\sqrt{nh_1}\left\{\hat{g}_1^+(u_1) - g_1(u_1) - \mu_1^* - \tfrac{1}{2} h_1^2 \mu_2(K) g_1''(u_1) + o\left(h_1^2\right)\right\} \to N(0, v^*(u_1)),$$

with $p_1, p_2, p_{1,2}$ *being the densities of* U_1, X_2 *and* $(U_1; X_2)$, *respectively*,

$$v^*(u_1) = p_1(u_1) \|K\|^2 E\left\{\sigma^2(X) \frac{W_2^2(X_2) p_2(X_2)}{p_{1,2}^2(U_1, X_2)} e_1^T \Sigma_1^{-1} \Sigma_2 \Sigma_1^{-1} e_1 \mid U_1 = u_1\right\},$$

$$\Sigma_1 = E\left\{\begin{pmatrix} 1 & X_3^T \\ X_3 & X_3 X_3^T \end{pmatrix} \middle| U_1, X_2\right\},$$

$$\Sigma_2 = \begin{pmatrix} 1 & X_3^T \\ X_3 & X_3 X_3^T \end{pmatrix}.$$

REMARK 11. If we apply the estimating procedure to each additive component of model (1.3), then the resulting estimators are asymptotically independent and normal.

Next, we estimate the parameter β. Let $\hat{\mu}^*$ and $\hat{\beta}^{**}$ minimize (4.6) with $\hat{g}_\alpha^*$ replaced by $\hat{g}_\alpha^+$. Then we can compute explicitly the asymptotic variance of $\hat{\beta}^{**}$ in a similar fashion to Theorem 4. Since the notation gets very complicated, we only state a simpler version of it.

THEOREM 6. *Under the assumptions of Theorem* 4, *we have*

$$\sqrt{n}\,(\hat{\beta}^{**} - \beta) \to N(0, B_3)$$

for some positive definite matrix B_3.

The proof of this theorem is similar to that of Theorem 4 and is omitted.

REMARK 12. Theorems 5 and 6 can be extended to the case that X_3 is continuous.

REMARK 13. For the case of one nonparametric component ($p = 1$) and continuous X_3, Speckman (1988) has shown that another method leads to an unbiased estimate of β. The approach of Speckman does not require undersmoothing [i.e., $h_{1\alpha} = o(n^{-1/4})$]. The estimate is based on the regression of $(I - M_S)Y$ onto $(I - M_S)X_3$, where M_S denotes a smoothing matrix. It is not clear to us how this approach generalizes to the case with more than one additive components. Efficient estimation of β for $p = 1$ has been considered in Bhattacharya and Zhao (1997).

4.3. *Exploring possible interactions*. Suppose one is interested in validating the additive model (1.2) by checking whether there is a nonnegligible interaction term such as $g_{12}(u_1, u_2)$. One can embed the additive model (1.2) into the model (1.4) or more generally model (1.1) with $p = 2$. Now, estimate the function $\hat{g}_{12}$ using our method. Plot $\hat{g}_{12}(\cdot; x_2)$ for a few different values of x_2. The parallelism of the plot suggests the additivity contributions of x_1 and x_2. This provides a quick and informal model diagnostic tool.

5. Simulations and an application. In a small simulation study we have compared the "indicator method" [see (2.4)] and the "linear approach" where the linear parametric part has been incorporated in the local linear smoothing [see (4.7)]. In our simulation and in the following data example we have not studied estimation of the optimal weight function W. First experience suggests that a practically working adaptation of this idea needs some further research.

We have generated 100 samples of 200 normal observations Y. Four covariates have been generated: U_1 and U_2 are normal with mean 0, variance 1 and covariance 0.4; Z_1 takes values 1, 2, 3 and 4 with probability 0.25, 0.35, 0.25 or 0.15, respectively; Z_2 takes values 0 or 1 with probability 0.2 or 0.8, respectively. The (conditional) variance of Y is $\{1 + (U_1^2 + U_2^2 + Z_1^2 + Z_2^2)^{1/2}\}/4$. The simulated regression function is $1.5 + g_1(u_1) + g_2(u_2) + \beta_1 z_1 - \beta_2 z_2$ with $g_1(u_1) = 1 - u_1^2$, $g_2(u_2) = \sin(-u_2)$, $\beta_1 = 0.3$ and $\beta_2 = -0.5$. In the estimation of the parametric components only observations have been used with $|U_1| \le 1.5$ and $|U_2| \le 1.5$; see (4.6). Bandwidths 0.3 and 0.4 have been used for the smoothing of the estimated or the nuisance nonparametric component, respectively. Table 1 shows the simulated MASE (i.e., the

TABLE 1
*Results from a small simulation study comparing the "linear approach" and the "indicator method." Two nonparametric additive components g_1 and g_2; two linear parameters β_1 and β_2; sample size $n = 200$**

		$\hat{g}_1$	$\hat{g}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{m}$
MASE	Indicator Method	0.1857 (0.0609)	0.1775 (0.518)	0.0096	0.0409	0.2647
	Linear Approach	0.2739 (0.1450)	0.3207 (0.1549)	0.0075	0.0393	0.5081

*In parentheses the MASE are given for the nonparametric components with summation region truncated by the 2.5% and 97.5% quantiles of the covariates.

squared error averaged over the design points). The values in parentheses are the MASE for the nonparametric components with the summation region truncated by the 2.5% and 97.5% quantiles of the covariates. These values have been added because they reflect better the behavior of the curve estimates in the middle region.

In this simulation the "indicator method" clearly shows a better performance. We conjecture that the "indicator method" may be outperformed by the "linear method" only in cases where the discrete variables take on a rather large number of different values. In the following data example we used the "indicator method."

Figure 1 contains the resulting plots from a study on the female labor supply in East Germany. A sample of 607 women with a job who live together with a partner were asked their weekly number Y of working hours. Furthermore, the following information was recorded: if the woman has children less than 16 years old (Z_1), the unemployment rate Z_2 in the "land" of the Federal Republic of Germany where she lives, the age U_1 of the woman, her wage per hour U_2, the "Treiman prestige index" of her job U_3 [see Treiman (1978)], her years U_4 of education (introduction of this covariate makes sense because of the strongly regulated system of education in the former East Germany), her rent or redemption U_5, and the monthly net income U_6 of her husband. A partial linear model for these data has been fitted. The fit has been chosen linearly in Z_1 and Z_2. The covariate Z_2 takes only five values. (There are five "lands" in the eastern part of Germany.) The other six additive components have been estimated nonparametrically. For this data set a constant weight function W has been used. Bandwidths 0.4 and 0.6 times the empirical standard deviation of the covariable have been used for the smoothing of the estimated or the nuisance nonparametric component, respectively. The resulting parametric estimates are $\beta_1 = -1.46$ and $\beta_2 = 0.52$. The resulting nonparametric fits can be found in the left frames of Figure 1. Dashed lines have been added for indicating the pointwise variance of the curve estimates. These lines differ from the curve estimates by 1.64

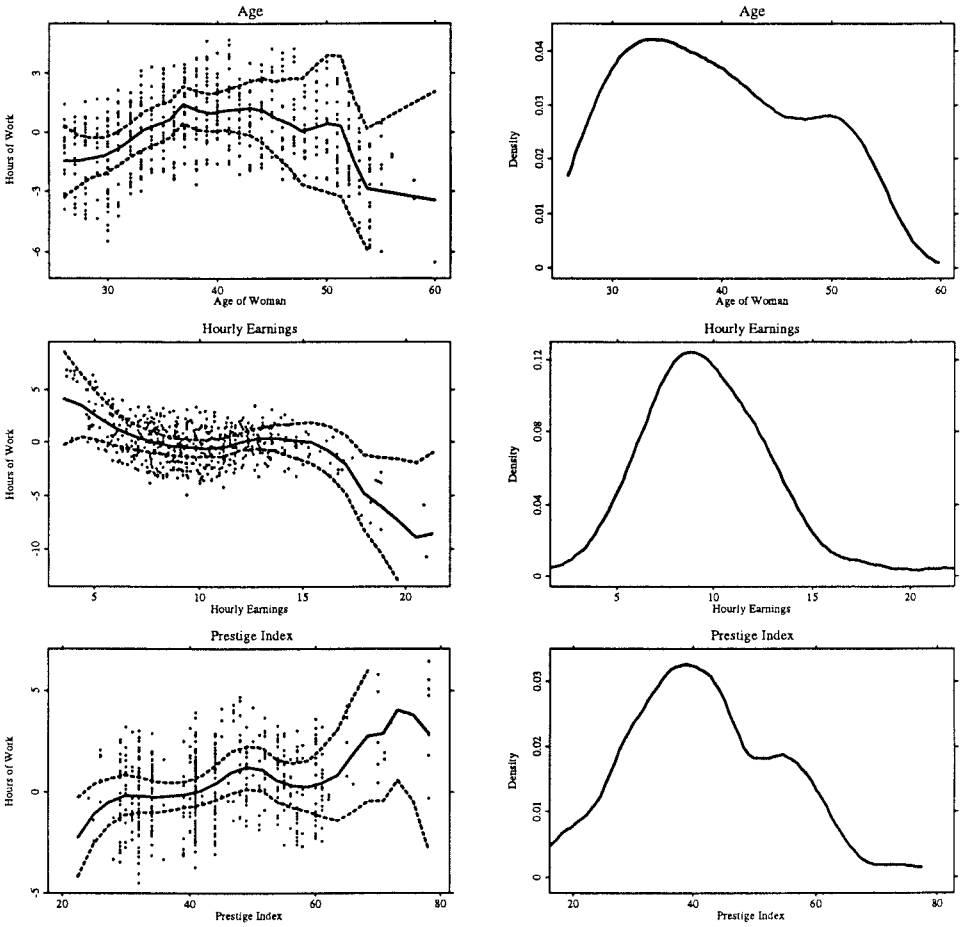

FIG. 1. *Female labor supply in East Germany. Left frames show nonparametric estimates of additive components with approximate* 90% *confidence intervals. Right frames give kernel density estimates of the covariates.*

times the (estimated) pointwise standard deviation of the curve estimates; that is, this corresponds to an approximate 90% confidence interval (without bias correction). The estimation of the pointwise standard deviation of the curve estimates has been done under the additional assumption that the conditional variance of the errors is constant. Note that all curve estimates at a fixed point are averages $\sum w_i Y_i$ of the observations Y_i. The variance of this estimate can be estimated by $\sum w_i^2 \hat{\sigma}_2$, where $\hat{\sigma}^2$ is the empirical variance of the residuals $\hat{\varepsilon} = Y - \hat{\mu} - \sum_j \hat{f}_j(U_j) - \sum_j \hat{\beta}_j Z_j$. Another estimate of the variance of $\sum w_i Y_i$ is $\sum w_i^2 \hat{\varepsilon}_i^2$. This estimate does not require the additional assumption that the conditional variance of the errors is constant. Plots of

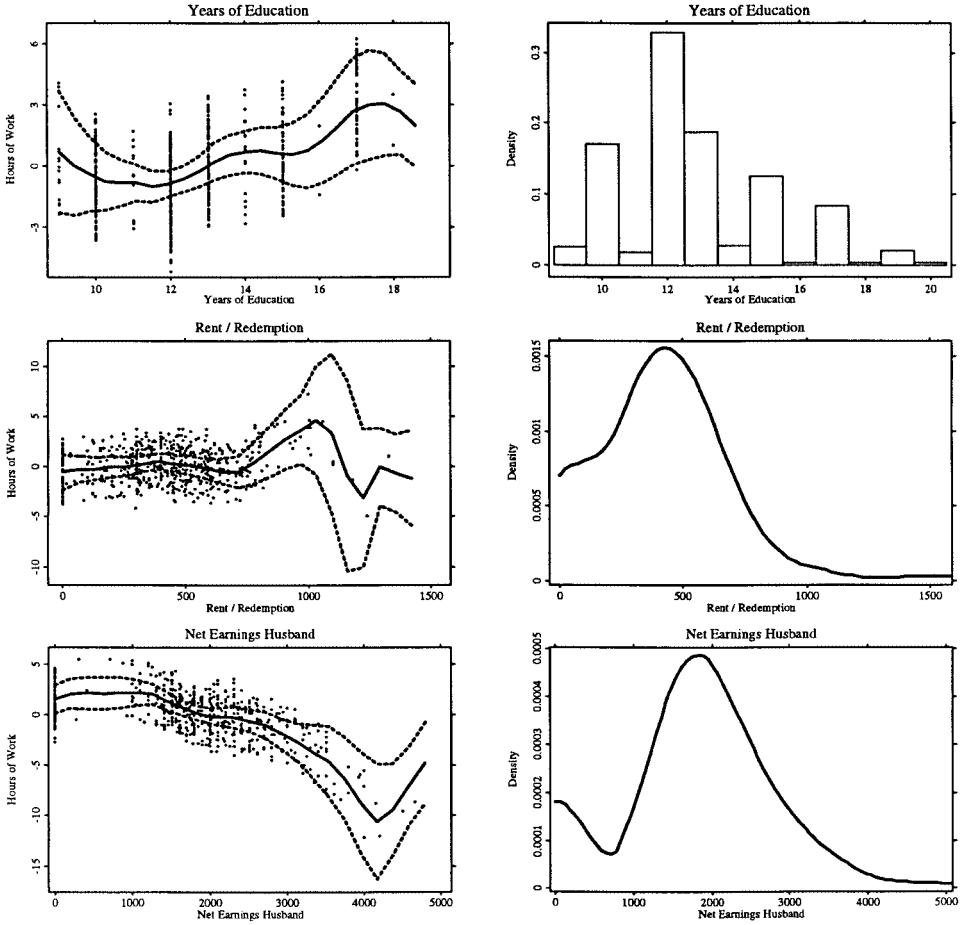

FIG. 1. *Continued*

pointwise confidence intervals based on this estimate are of similar size but of rougher shape. They are not shown here. In each plot of Figure 1 the covariable has been plotted against the estimated function plus the logarithm of the residual [i.e., $\hat{f}_\alpha(U_\alpha) + \text{sgn}(\hat{\varepsilon})\log|\hat{\varepsilon}|$; the logarithmic transform has been used to show all data]. The right frames show the density estimates of the covariates.

The plots show some clear nonlinearities. In particular, one sees a flat part in the lower range for rent and prestige index and in the middle range of hourly earnings, whereas the relation is monotone elsewhere. The results quantify the extent to which each variable affects the female labor supply. Using the dynamic ranges of the plots as a criterion to assess the practical

importance of a variable, the key factors that affect the labor supply are hourly earnings U_2 and monthly net income of husbands U_6. Slightly less influential covariates are age of the woman U_1 and prestige of the job U_2. Table 2 shows the results of a parametric least-squares analysis.

The covariates $U_1^2 = (\text{AGE W.})^2$ and $U_2^2 = (\text{WAGE P. H.})^2$ have been introduced in the parametric model. The presence of these quadratic terms is highly significant. The introduction of U_1^2 is motivated partially by the shape of the nonparametric estimate of g_1. There is no significant change in the values of the parameters β_1 and β_2. Otherwise, there are some differences between the parametric and the semiparametric analysis. Clearly, the piecewise linear shape of g_2, g_3, and g_5 cannot be recovered in the parametric model. For g_5 the sign of the estimated parameter agrees with the slope of the nonparametric estimate in the upper part. Note that for g_2 the parametric analysis with covariates U_2 and U_2^2 differs strongly for the upper part of g_2. At the boundaries of the functions g_4, g_5 and g_6 we see some differences between the parametric analysis and the semiparametric analysis. Clearly, the boundary behavior of the nonparametric estimates depends on a relatively small fraction of the observations. For example, the monotone decreasing part at the beginning of g_4, is caused by only 15 women with 9 years of education and an introduction of a covariate U_4^2 in the parametric analysis is not significant.

It seems to be difficult to verify the data analytic findings of a semiparametric analysis. A first step is to consider test statistics which are based on the comparison of parametric and nonparametric fits; see, for instance,

TABLE 2
Female labor supply in East Germany. Results of an ordinary least-squares analysis

Source	Sum of squares	Degrees of freedom	Mean square	F-ratio
Regression	6526.3	10	652.6	9.24
Residual	42,101.1	596	70.6	
R squared	= 13.4%	R squared (adjusted)	= 12.0%	

Variable	**Estimate**	**Standard error**	***t*-value**	**Probability > [*t*]**
CONSTANT	1.36	8.95	0.15	0.8797
CHILD	−2.63	1.09	−2.41	0.0163
UNEMPLOYMENT	0.48	0.22	2.13	0.0333
AGE W.	1.63	0.43	3.75	0.0002
(AGE W.)^2	−0.021	0.0054	−3.82	0.0001
WAGE P. H.	−1.07	0.18	−6.11	≤ 0.0001
(WAGE P. H.)^2	0.0017	0.0033	4.96	≤ 0.0001
PRESTIGE	0.13	0.034	3.69	0.0002
YEARS EDUC.	0.66	0.19	3.58	0.0004
RENT/RED.	0.0018	0.0012	1.56	0.1198
NET INC. H.	−0.0016	0.0003	−4.75	≤ 0.0001

Härdle and Mammen (1993) and Härdle, Mammen and Müller (1995). The second paper discusses also extensions to generalized regression.

6. Conditions and proofs.

CONDITION A. (i) We suppose that the functions W and f_2 are bounded on the support S of W. The weight function $W(x_2, x_3)$ is uniformly continous with respect to x_2.

(ii) The kernel functions K and L are symmetric and have bounded supports. Furthermore, L is an order-d kernel.

(iii) The support of the discrete variable X_3 is finite and

$$\inf_{\substack{u_1 \in x_1 \pm \delta \\ (x_2, x_3) \in S}} p_3(x_3)\, p_{1,2\,|\,3}(u_1, x_2 \mid x_3) > 0 \quad \text{for some } \delta > 0.$$

For u_1 in a neighborhood of x_1 and for (u_2, u_3) in S, the conditional density $p_{1,2\,|\,3}(u_1, u_2 \mid u_3)$ has bounded partial derivatives up to order 2 with respect to u_1 and up to order d with respect to u_2.

(iv) f_1 has a bounded second derivative in a neighborhood of x_1 and $f(x_2, x_3)$ has a bounded dth-order derivative with respect to x_2.

(v) $E\varepsilon^4$ is finite and $\sigma^2(x) = E(\varepsilon^2 \mid X = x)$ is continuous, where $\varepsilon = Y - E(Y \mid X)$. Furthermore, for a $\delta > 0$, the conditional absolute moment $E(|\varepsilon|^{2+\delta} | X_1 = u_1)$ is bounded for u_1 in a neighborhood of x_1.

(vi) $nh_1^p h_2^{2q}/\log^2 n \to \infty$ and $h_1^4 \log n / h_2^q \to 0$.

CONDITION B. (i) The functions $g_{-\alpha}$ and W_α are bounded on the support S_α of W_α. The weight function W_α is uniformly continuous.

(ii) The same as Condition A(ii).

(iii) $\inf p(u_1, \ldots, u_p) > 0$, where the infimum runs over $u_\alpha \in x_\alpha \pm \delta$ and $(u_1, \ldots, u_{\alpha-1}, u_{\alpha+1}, \ldots, u_p) \in S_\alpha$. For u_1 in a neighborhood of x_1 and for $(u_1, \ldots, u_{\alpha-1}, u_{\alpha+1}, \ldots, u_p) \in S_\alpha$, the density p has bounded partial derivatives up to order 2 with respect to u_α and up to order d with respect to u_β, $\beta \neq \alpha$.

(iv) g_α has bounded and continous derivatives up to order 2 and g_β, $\beta \neq \alpha$, have bounded and continous derivatives up to order d.

(v) The same as Condition A(v).

(vi) $nh_1 h_2^{2(p-1)}/\log^2 n \to \infty$ and $h_1^4 \log n / h_2^{p-1} \to 0$.

PROOF OF THEOREM 1. Let $x^i = (x_1, X_{2i}, X_{3i})$ and let E_i denote the conditional expectation given $X_i = (X_{1i}, X_{2i}, X_{3i})$. Denote by $p(x) = p_3(x_3) p_{1,2\,|\,3}(x_1, x_2 \mid x_3)$. Then, by (2.1) and Condition A(i), we have

$$n^{-1} \sum_{i=1}^{n} m(x^i) W(X_{2i}, X_{3i}) = f_1^*(x_1) + O_p(n^{-1/2}). \tag{6.1}$$

Thus,

$$\hat{f}_1^*(x_1) - f_1^*(x_1) = n^{-1}\sum_{i=1}^{n}\{\hat{m}(x^i) - m(x^i)\}W(X_{2i}, X_{3i}) + O_p(n^{-1/2}). \tag{6.2}$$

Let $\hat{r}_{ij} = m(X_j) - m(x^i) - f_1'(x_1)^T(X_{1j} - x_1)$ and let $\hat{r}_i$ be the resulting $n \times 1$ vector. Then, by (1.1) and the definition of K_n, it follows that

$$\begin{aligned}&\hat{m}(x^i) - m(x^i)\\ &\quad = e_1^T S_n^{-1}(x^i)\begin{pmatrix} 1 & \cdots & 1 \\ X_{11} - x_1 & \cdots & X_{1n} - x_1 \end{pmatrix} A(x^i)(\hat{r}_i + \tilde{\varepsilon}),\end{aligned} \tag{6.3}$$

where $A(x)$ is a diagonal matrix with diagonal elements $A_i(x) = K_{h_1}(X_{1i} - x_1)L_{h_2}(X_{2i} - x_2)I\{X_{3i} = x_3\}$ and $\tilde{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$ with $\varepsilon_i = Y_i - m(X_i)$. Let $H = \operatorname{diag}(1, h_1^{-1}, \ldots, h_1^{-1})$ be a $(p+1) \times (p+1)$ diagonal matrix and $a_n = \{\log n/(nh_1^p h_2^q)\}^{1/2}$. Then, owing to the uniform convergence of the kernel density estimator [cf. Stone (1993)], we have

$$\begin{aligned}&n^{-1}HS_n(x)H\\ &\quad = n^{-1}\sum_{i=1}^{n} A_i(x)\begin{pmatrix} 1 \\ (X_{1i} - x_1)/h_1 \end{pmatrix}\begin{pmatrix} 1 \\ (X_{1i} - x_1)/h_1 \end{pmatrix}^T\\ &\quad = EA_i(x)\begin{pmatrix} 1 \\ (X_{1i} - x_1)/h_1 \end{pmatrix}\begin{pmatrix} 1 \\ (X_{1i} - x_1)/h_1 \end{pmatrix}^T + O_p(a_n)\\ &\quad = \begin{pmatrix} p(x) & h_1 p^{(1,0)}(x)^T\mu_2(K) \\ h_1\mu_2(K)p^{(1,0)}(x) & p(x)\mu_2(K) \end{pmatrix} + O_p(c_n)\\ &\quad = \begin{pmatrix} p(x) & 0 \\ 0 & p(x)\mu_2(K) \end{pmatrix} + o_p(c_n),\end{aligned} \tag{6.4}$$

where $c_n = h_1^2 + h_2^d + a_n$ and where $p^{(1,0)}$ denotes the vector of partial derivatives of p with respect to x_1. Now note that

$$\begin{aligned}&\begin{pmatrix} p(x) & h_1 p^{(1,0)}(x)^T\mu_2(K) \\ h_1\mu_2(K)p^{(1,0)}(x) & p(x)\mu_2(K) \end{pmatrix}^{-1}\\ &\quad = \begin{pmatrix} p(x) & 0 \\ 0 & p(x)\mu_2(K) \end{pmatrix}^{-1}\\ &\qquad + \frac{h_1}{p(x)}\begin{pmatrix} 0 & p^{(1,0)}(x)^T\mu_2(K) \\ p^{(1,0)}(x)\mu_2(K) & 0 \end{pmatrix} + O_p(h_1^2).\end{aligned}$$

A similar argument to the above leads to the following uniform results:

$$\begin{aligned} & n^{-1}H\begin{pmatrix} 1 & \cdots & 1 \\ X_{11}-x_1 & \cdots & X_{1n}-x_1 \end{pmatrix} A(x^i)\hat{r}_i \\ &= n^{-1}\sum_{j=1}^{n} A_j(x^i)\hat{r}_{ij}\begin{pmatrix} 1 \\ (X_{1j}-x_1)/h_1 \end{pmatrix} \\ &= E_i A_j(X^i)\hat{r}_{ij}\begin{pmatrix} 1 \\ (X_{1j}-x_1)/h_1 \end{pmatrix} + O_p(a_n) \\ &= O_p(c_n), \end{aligned}$$

where in the third expression j is an arbitrary index with $j \neq i$ and

$$\begin{aligned} & n^{-1}H\begin{pmatrix} 1 & \cdots & 1 \\ X_{11}-x_1 & \cdots & X_{1n}-x_1 \end{pmatrix} A(x^i)\tilde{\varepsilon} \\ &= n^{-1}\sum_{j=1}^{n} A_j(x^i)\varepsilon_j\begin{pmatrix} 1 \\ (X_{1j}-x_1)/h_1 \end{pmatrix} \\ &= O_p(a_n). \end{aligned}$$

Substituting all of the above expressions into (6.3), after some algebra, we obtain

$$\begin{aligned} & \hat{m}(x^i) - m(x^i) \\ &= e_1^T\left[\begin{pmatrix} p(x^i) & 0 \\ 0 & p(x^i)\mu_2(K) \end{pmatrix}^{-1}\right. \\ &\qquad \left. + \frac{h_1}{p(x^i)}\begin{pmatrix} 0 & p^{(1,0)}(x^i)^T\mu_2(K) \\ p^{(1,0)}(x^i)\mu_2(K) & 0 \end{pmatrix} + O_p(c_n)\right] \\ &\qquad \times n^{-1}\sum_{j=1}^{n} A_j(x^i)\big(\hat{r}_{ij}+\varepsilon_j\big)\begin{pmatrix} 2 \\ (X_{1j}-x_1)/h_1 \end{pmatrix} \\ &= n^{-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} A_j(x^i)\big(\hat{r}_{ij}+\varepsilon_j\big)/p(x^i) \\ &\qquad + n^{-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} A_j(x^i)\big(\hat{r}_{ij}+\varepsilon_j\big)p^{-1}(x^i)p^{(1,0)}(x^i)^T\mu_2(K)(X_{1j}-x_1) \\ &\qquad + O_p\big(c_n^2\big). \end{aligned}$$

Clearly, the second term will be smaller than the first one by an order of $O(h_1)$. When the above O_p-term is averaged in (6.2), it is still of the order

$$O_p\big(c_n^2\big) = o_p\big((nh_1^p)^{-1/2}\big)$$

by the conditions on the bandwidths. Furthermore, by calculation of the first two moments one can show that

$$n^{-2}\sum_{j\neq i} W(X_{2i}, X_{3i})\, A_j(x^i)\hat{r}_{ij} p^{-2}(x^i)\, p^{(1,0)}(x^i)^T \mu_2(K)(X_{1j} - x_1)$$
$$= o(h_1^2) + O(h_2^d) + o_p\big((nh_1^p)^{-1/2}\big)$$

and

$$n^{-2}\sum_{j\neq i} W(X_{2i}, X_{3i})\, A_j(x^i)\varepsilon_j p^{-2}(x^i)\, p^{(1,0)}(x^i)^T \mu_2(K)(X_{1j} - x_1)$$
$$= o_p\big((nh_1^p)^{-1/2}\big).$$

In other words, the approximation error from (6.4) is negligible.

Note that, for $j \neq i$,

$$E_i A_j(x^i)\hat{r}_{ij} = \tfrac{1}{2}h_1^2 \operatorname{tr}\{f_1''(x_1)\mu_2(K)\}p(x^i) + o(h_1^2) + O(h_2^d).$$

Let $\tilde{r}_{ij} = A_j(x^i)\hat{r}_{ij} - E_i A_j(x^i)\hat{r}_{ij}$ for $j \neq i$ and $\tilde{r}_{ij} = 0$ for $j = i$. Thus, by (6.2), we have

$$\hat{f}_1^*(x_1) - f_1^*(x_1) = \tfrac{1}{2}h_1^2 \operatorname{tr}\{f_1''(x_1)\mu_2(K)\} + o(h_1^2) + T_{n,1} + T_{n,2} + o_p\big\{(nh_1^p)^{-1/2}\big\}, \tag{6.5}$$

where

$$T_{n,1} = n^{-1}\sum_{j\neq i} \varepsilon_j K_{h_1}(X_{1j} - x_1)\Gamma(X_{2i}, X_{3i}) L_{h_2}(X_{2j} - X_{2i}) I\{X_{3j} = X_{3i}\}$$

and

$$T_{n,2} = n^{-2}\sum_{j\neq i} \Gamma(X_{2i}, X_{3i})\tilde{r}_{ij},$$

with

$$\Gamma(X_{2i}, X_{3i}) = W(X_{2i}, X_{3i})/p(x^i).$$

We will show that with $\varepsilon_j^* = G(X_{2j}, X_{3j})\varepsilon_j$, $G(X_{2j}, X_{3j}) = \Gamma(X_{2j}, X_{3j})p_{2,3}(X_{2j}, X_{3j})$,

$$T_{n,1} = n^{-1}\sum_{j=1}^{n} K_{h_1}(X_{1j} - x_i)\varepsilon_j^* + o_p\big((nh_1^p)^{-1/2}\big) \tag{6.6}$$

and

$$T_{n,2} = o_p\big((nh_1^p)^{-1/2}\big). \tag{6.7}$$

Combination of (6.5)–(6.7) leads to

$$\hat{f}_1^*(x_1) - f_1^*(x_1) = \tfrac{1}{2}h_1^2 \operatorname{tr}\{f_1''(x_1)\mu_2(K)\} + n^{-1}\sum_{j=1}^{n} \varepsilon_j^* K_{h_1}(X_{1j} - x_1) + o_p\big\{(nh_1^p)^{-1/2}\big\}. \tag{6.8}$$

It is easy to show that

$$\sqrt{nh_1^p}\, n^{-1} \sum_{j=1}^{n} \varepsilon_j^* K_{h_1}(X_{1j} - x_1) \to \mathcal{N}(0, v(x_1)) \tag{6.9}$$

by checking the Lyapounov condition. By using (6.8) and (6.9), we establish Theorem 1. It remains to verify (6.7) and (6.8).

PROOF OF (6.6). Let

$$V_{i,j} = \Gamma(X_{2i}, X_{3i}) L_{h_2}(X_{2j} - X_{2i}) I\{X_{3i} = X_{3j}\} - p_3(X_{3j}) p_{2|3}(X_{2j} \mid X_{3j}) \Gamma(X_{2j}, X_{3j}).$$

Note that, for $i \neq j$,

$$E_j V_{i,j} = p_3(X_{3j}) \int \Gamma(x_2, X_{3j}) L_{h_2}(x_2 - X_{2j}) p_{2|3}(x_2 \mid X_{3j})\, dx_2 - p_3(X_{3j}) p_{2|3}(X_{2j} \mid X_{3j}) \Gamma(X_{2j}, X_{3j}).$$

Thus,

$$|E_j V_{i,j}| \leq \int \big| \Gamma(X_{2j} + h_2 u, X_{3j}) p_{2|3}(X_{2j} + h_2 u \mid X_{3j}) - \Gamma(X_{2j}, X_{3j}) p_{2|3}(X_{2j} \mid X_{3j}) \big| \, |L(u)|\, du \to 0.$$

Note also that the difference between the left-hand side of (6.6) and the main term on the right-hand side of (6.6) can be expressed as

$$D_{n,1} = n^{-2} \sum_{j \neq i} \varepsilon_j K_{h_1}(X_{1j} - x_1) V_{i,j}.$$

To prove (6.6), it suffices to show

$$ED_{n,1}^2 = o\big((nh_1^p)^{-1}\big).$$

It follows from direct expansion that

$$ED_{n,1}^2 = n^{-4} \sum_{i \neq j;\, k \neq l} E\varepsilon_j K_{h_1}(X_{1j} - x_1) V_{i,j} \varepsilon_l K_{h_1}(X_{1l} - x_1) V_{k,l}.$$

Because of $E\{\varepsilon_j \mid X_j\} = 0$ we have

$$ED_{n,1}^2 = n^{-4} \sum_{i \neq j;\, k \neq j} E\varepsilon_j^2 K_{h_1}(X_{1j} - x_1) V_{i,j} \varepsilon_l K_{h_1}(X_{1l} - x_1) V_{k,j}.$$

For $i = k$ the order of summands on the right-hand side is at most $O(h_1^{-p} h_2^{-q})$. Because of $n^{-2} h_1^{-p} h_2^{-q} = o(n^{-1})$, we have

$$\begin{aligned} ED_{n,1}^2 &= n^{-4} \sum_{i \neq j \neq k \neq i} E\varepsilon_j^2 K_{h_1}^2(X_{1j} - x_1) V_{i,j} V_{k,j} + o(n^{-1}) \\ &= n^{-4} \sum_{i \neq j \neq k \neq i} E\varepsilon_j^2 K_{h_1}^2(X_{1j} - x_1) E_j V_{i,j} E_j V_{k,j} + o(n^{-1}) \\ &= o\big(n^{-1} h_1^{-p}\big). \end{aligned}$$

PROOF OF (6.7). The claim follows from $ET_{n,2}^2 = o((nh_1^p)^{-1})$. Note that $E_i\tilde{r}_{i,j} = 0$. Therefore, for the calculation of $ET_{n,2}^2$ we need only consider terms of the form

$$n^{-4}E\Gamma(X_{2i}, X_{3i})\tilde{r}_{i,j}\Gamma(X_{2k}, X_{3k})\tilde{r}_{k,l},$$

where $i \neq j, k \neq l$, $j \in \{k, l\}$ and $l \in \{i, j\}$. It is easy to bound the summands for two different indices. For three different indices we have $j = l$ and $i \neq j \neq k \neq i$. Note now that for this case

$$E\Gamma(X_{2i}, X_{3i})\tilde{r}_{i,j}\Gamma(X_{2k}, X_{3k})\tilde{r}_{k,j} = O\big(h_1^{-p}\big[h_1^4 + h_2^2\big]\big).$$

Here we have used that the random variables $\hat{r}_{i,j}$ are always bounded by a constant which is of order $O(h_1^2 + h_2)$. Thus,

$$ET_{n,2}^2 = O\big(n^{-1}h_1^{-p}\big[h_1^4 + h_2^2\big]\big) = o\big((nh_1^p)^{-1}\big),$$

verifying (6.7).

PROOF OF THEOREM 3. By (6.5)–(6.7), each component of $\hat{g}_\alpha^*(u_\alpha)$ has the following stochastic representation:

$$\begin{aligned}
&\hat{g}_\alpha^*(u_\alpha) - g_\alpha(u_\alpha) - \mu_{1\alpha} \\
&\quad = \tfrac{1}{2}h_{1\alpha}^2\mu_2(K)g_\alpha''(u_\alpha) + o\big(h_{1\alpha}^2\big) \\
&\qquad + n^{-1}\sum_{j=1}^{n} K_{h_{1\alpha}}(U_{\alpha j} - U_\alpha)G_\alpha(U_{-\alpha j})\varepsilon_j + o_p\big((nh_{1\alpha})^{-1/2}\big),
\end{aligned} \tag{6.10}$$

where

$$G_\alpha(U_{-\alpha j}) = \frac{W_\alpha(U_{-\alpha j})p_{-\alpha}(U_{-\alpha j})}{p(U_\alpha^j)},$$

with $U_\alpha^j = (U_{1j}, \ldots, U_{\alpha-1,j}, u_\alpha, U_{\alpha+1,j}, \ldots, U_{pj})$. For $\alpha \neq \beta$, the covariance for the stochastic terms in (6.10) is

$$\begin{aligned}
&\operatorname{cov}\left(n^{-1}\sum_{j=1}^{n} K_{h_{1\alpha}}(U_{\alpha j} - u_\alpha)G_\alpha(U_{-\alpha j})\varepsilon_j, n^{-1}\sum_{j=1}^{n} K_{h_{1\beta}}(U_{\beta j} - u_\beta)G_\beta(U_{-\beta j})\varepsilon_j\right) \\
&\quad = n^{-1}E\big[K_{h_{1\alpha}}(U_\alpha - u_\alpha)G_\alpha(U_{-\alpha})K_{h_{1\beta}}(U_\beta - u_\beta)G_\beta(U_{-\beta})\varepsilon^2\big] \\
&\quad = O(n^{-1}) = o\big((nh_{1\alpha})^{-1/2}(nh_{1\beta})^{-1/2}\big).
\end{aligned}$$

Therefore, the asymptotic covariance should be 0.

PROOF OF THEOREM 4. We only outline the key steps of the proof. Proceeding as in the proof of Theorem 1, one shows first that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_i I(X^i \in A)\big[\hat{g}_\alpha^*(U_{\alpha i}) - g_\alpha(U_{\alpha i}) - \mu_{1\alpha}\big] &= \frac{1}{n}\sum_{i=1}^{n} Z_i I(X^i \in A)\frac{1}{n}\sum_{j\neq k} K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i}) \\ &\quad \times L_{h_{2\alpha}}(U_{-\alpha j} - U_{-\alpha k}) I(X_{3j} = X_{3k}) \\ &\quad \times \big[m(X^j) - m(U_{\alpha i}, U_{-\alpha k}, X_{3k}) \\ &\qquad -g'_\alpha(U_{\alpha i})(U_{\alpha j} - U_{\alpha i}) + \varepsilon_k\big] \\ &\quad \times \frac{1}{p(U_{\alpha i}, U_{-\alpha k}, X_{3k})} + O_P\big(c_n^2\big) + o_p\big(n^{-1/2}\big),\end{aligned}$$

where now $c_n = h_{1\alpha}^2 + h_{2\alpha}^d + \{\log n/(nh_{1\alpha}h_{2\alpha}^{p-1})\}^{1/2}$. Note that under our assumptions we have $c_n^2 = o(n^{-1/2})$. By considering the first two moments of the difference, one can show that (see also the asymptotic treatment of $T_{n,1}$ and $T_{n,2}$ in the proof of Theorem 1)

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_i I(X^i \in A)\big[\hat{g}_\alpha^*(U_{\alpha i}) - g_\alpha(U_{\alpha i}) - \mu_{1\alpha}\big] &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i I(X^i \in A) K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i}) \\ &\quad \times G_\alpha(U_{\alpha i}, U_{-\alpha j}, X_{3,j})\varepsilon_j + o_p\big(n^{-1/2}\big),\end{aligned} \tag{6.11}$$

where

$$G_\alpha(U_{\alpha i}, U_{-\alpha j}, X_{3,j}) = \frac{W_\alpha(U_{-\alpha j}, X_{3,j}) p_{-\alpha,3}(U_{-\alpha j}, X_{3,j})}{p(U_{\alpha i}, U_{-\alpha j}, X_{3,j})}.$$

Thus, the main term of $\hat{\beta}^*$ is

$$\hat{\beta}^* = \beta^* + \big(\tilde{Z}^T \Delta \tilde{Z}\big)^{-1} \tilde{Z}^T \Delta \bar{\varepsilon} + o_p\big(n^{-1/2}\big), \tag{6.12}$$

where the ith element of $\bar{\varepsilon}$ is

$$\begin{aligned}\bar{\varepsilon}_i = \varepsilon_i - n^{-1}\sum_{\alpha=1}^{p}\sum_{j=1}^{n} K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i}) \\ \times G_\alpha(U_{\alpha i}, U_{-\alpha j}, X_{3,j})\varepsilon_j - \sum_{\alpha=1}^{p} \varepsilon_{\alpha i}^*.\end{aligned} \tag{6.13}$$

Obviously, by the law of large numbers,

$$n^{-1}\tilde{Z}^T \Delta \tilde{Z} = E\big[I(X \in)ZZ^T\big] + o_p(1) = B_1 + o_p(1). \tag{6.14}$$

Let $\varepsilon^*_{\alpha i} = g_\alpha(U_{\alpha i})a_{n,\alpha} + b_{n,\alpha}$ be the approximation error in (6.1), where

$$a_{n,\alpha} = n^{-1}\sum_{j=1}^{n} W_\alpha(U_{-\alpha j}, X_{3j}) - 1$$

and

$$b_{n,\alpha} = n^{-1}\sum_{j=1}^{n}\left\{g_{-\alpha}(U_{-\alpha}) + X_{3j}^T\beta\right\}W_\alpha(U_{-\alpha j}, X_{3j}) - E\left[\left\{g_{-\alpha}(U_{-\alpha j}) + X_2^T\beta\right\}W_\alpha(U_{-\alpha}, X_3)\right].$$

We need only consider the term

$$\begin{aligned} n^{-1}\tilde{Z}^T\Delta\bar{\varepsilon} &= n^{-1}\sum_{i=1}^{n} Z_i\bar{\varepsilon}_i I\{X^i\in A\} \\ &= n^{-1}\sum_{i=1}^{n} Z_i\varepsilon_i I\{X^i\in A\} \\ &\quad - n^{-1}\sum_{\alpha=1}^{p}\sum_{i=1}^{n} Z_i I\{X^i\in A\}\left(g_\alpha(U_{\alpha i})a_{n,\alpha} + b_{n,\alpha}\right) \\ &\quad - n^{-1}\sum_{i=1}^{n}\varepsilon_i n^{-1}\sum_{\alpha=1}^{p}\sum_{j=1}^{n} G_\alpha(U_{\alpha j}, U_{-\alpha i}, X_{3,i}) \\ &\qquad\qquad \times K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i})I\{X^j\in A\}. \end{aligned} \tag{6.15}$$

By using the same argument as in the proof of (6.6), we can show that

$$\begin{aligned} &n^{-1}\sum_{i=1}^{n}\varepsilon_i n^{-1}\sum_{j=1}^{n} G_\alpha(U_{\alpha j}, U_{-\alpha i}, X_{3,i})K_{h_{1\alpha}}(U_{\alpha j} - U_{\alpha i})Z_j I\{X^j\in A\} \\ &= n^{-1}\sum_{i=1}^{n}\varepsilon_i G_\alpha(U_{\alpha i}, U_{-\alpha i}, X_{3,i}) \\ &\qquad \times E\left\{Z_i I(X^i\in A)\mid U_{\alpha i}\right\}p_\alpha(U_{\alpha i}) + o_p(n^{-1/2}). \end{aligned} \tag{6.16}$$

Let $Z_{i,A} = Z_i I\{X^i\in A\} - \sum_{\alpha=1}^{p} G_\alpha(U_{\alpha i}, U_{-\alpha i}, X_{3,i})E\{Z_i I(X^i\in A)\mid U_{\alpha i}\}p_\alpha(U_{\alpha i})$. Then, by combining (6.15) and (6.16), we obtain

$$\begin{aligned} n^{-1/2}\tilde{Z}^T\Delta\bar{\varepsilon} &= n^{-1/2}\sum_{i=1}^{n} Z_{i,A}\varepsilon_i \\ &\quad - n^{-1/2}\sum_{\alpha=1}^{p} E\left[Z I\{X\in A\}g_\alpha(U_\alpha)\right]a_{n,\alpha} \\ &\quad + E\left[Z I\{X\in A\}\right]b_{n,\alpha} + o_p(1) \\ &\to N\left(0, E\left[\varepsilon_i^2 Z_{i,A}Z_{i,A}^T\right] + \operatorname{var}\left(\sum_{\alpha=1}^{p} V_\alpha\right)\right). \end{aligned} \tag{6.17}$$

By conditioning on X^i, one can easily see that the covariance matrix in (6.17) is B_2. Combination of (6.12), (6.14) and (6.17) shows the statement of Theorem 4.

PROOF OF THEOREM 5. The main ideas of the proof are the same as those of Theorem 1. Thus, we only indicate the main steps. Let $x^i = (u_1, X_{2i})$. Then we have

$$n^{-1}\sum_{i=1}^{n} m(x^i)W(X_{2i}) = g_1^+(u_1) + O_p(n^{-1/2}) \tag{6.18}$$

and

$$\hat{g}_1^+(u_1) - g_1^+(u_1) = n^{-1}\sum_{i=1}^{n}\{\hat{m}^*(x^i) - m(x^i)\}W(X_{2i}) + O_p(n^{-1/2}). \tag{6.19}$$

Set $A_j(u) = K_{h_1}(U_{1j} - u_1)L_{h_2}(X_{2j} - x_2)$. Let X be the design matrix of (4.7) and $A(u) = \operatorname{diag}(A_1(u), \ldots, A_n(u))$ be the corresponding weight matrix.

Denote by

$$\hat{r}_{ij} = g_1(U_{1j}) - g_1(u_1) - g_1'(u_1)(U_{1j} - u_1) + f_2(X_{2j}) - f_2(X_{2i}),$$

where $f_2(x_2) = g_2(u_2) + \cdots + g_p(u_p)$. Let $\hat{r}_i$ be the resulting $(n \times 1)$ vector. Then

$$\hat{g}^*(x^i) - g(x^i) = e_1^T S_n^{-1}(x^i)X^T A(x^i)\big(\hat{r}_i + \tilde{\varepsilon}\big), \tag{6.20}$$

where $S_n(x) = X^T A(x^i)X$. For $u = (u_1, x_2)$ let

$$S(u) = E\left\{\begin{pmatrix} 1 & 0 & X_3^T \\ 0 & \mu_2(K) & 0 \\ X_3 & 0 & X_3X_3^T \end{pmatrix}\middle| U_1 = u_1,\ X_2 = x_2\right\}$$

and

$$H = \begin{pmatrix} 1 & & & & \\ & h_1^{-1} & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}.$$

With the same ideas as in the proof of Theorem 1, one gets an expansion of $[n^{-1}HS_n(u)H]^{-1}$ up to error terms of order $O_P(c_n)$ where, as in the proof of Theorem 4, $c_n = h_{1\alpha}^2 + h_{2\alpha}^d + \{\log n/(nh_{1\alpha}h_{2\alpha}^{p-1})\}^{1/2}$. In particular, we have that

$$n^{-1}HS_n(u)H = p(u)S(u) + o_p(1) \tag{6.21}$$

uniformly in u. Direct calculation yields

$$n^{-1}E_iHX^TA(x^i)\hat{r}_i = \big\{\tfrac{1}{2}h_1^2 g_1''(u_1)\mu_2(K) + o(h_1^2) + O(h_2^d)\big\}p(x^i)\begin{pmatrix} 1 \\ 0 \\ E(X_{3i} \mid X^i) \end{pmatrix}. \tag{6.22}$$

Note now that

$$S(x^i)\begin{pmatrix}1\\0\\0\end{pmatrix} = \begin{pmatrix}1\\0\\E(X_{3i}\mid x^i)\end{pmatrix}.$$

Therefore,

$$e_1^T S^{-1}(x^i)\begin{pmatrix}1\\0\\E(X_{3i}\mid x^i)\end{pmatrix} = 1. \tag{6.23}$$

Substituting the higher-order expansion of $[n^{-1}HS_n(u)H]^{-1}$ and (6.22) into (6.20), we obtain with (6.23) that

$$\begin{aligned}\hat{g}^*(x^i) - g(x^i) &= \tfrac{1}{2}h_1^2 g_1''(u_1)\mu_2(K) + o(h_1^2) + O(h_2^d)\\ &\quad + n^{-1}\sum_{j=1}^n p(x^i)^{-1} e_1^T S^{-1}(x^i) A_j(x^i)\begin{pmatrix}1\\(U_{j1}-u_1)/h_1\\X_{3j}\end{pmatrix}\varepsilon_j\\ &\quad + n^{-1}\sum_{j=1}^n p(x^i)^{-1} e_1^T S^{-1}(x^i)\tilde{r}_{ij} + O_p(n^{-1/2}),\end{aligned} \tag{6.24}$$

where

$$\tilde{r}_{ij} = A_j(x^i)\hat{r}_{ij}\begin{pmatrix}1\\(U_{1j}-u_1)/h_1\\X_{3j}\end{pmatrix} - E_i A_j(x^i)\hat{r}_{ij}\begin{pmatrix}1\\(U_{1j}-u_1)/h_1\\X_{3j}\end{pmatrix}.$$

Note that again we obtain that the expansion of the estimate depends only on the first-order approximation (6.21) $n^{-1}HS_n(u)H$.

Using the same argument as in the proof of Theorem 1, the average of the last term in (6.24) over i is of order $o_p(n^{-1/2})$. Thus, by (6.19) and (6.24), we have

$$\begin{aligned}\hat{g}_1^+(u_1) - g_1^+(u_1) &= \tfrac{1}{2}h_1^2 g_1''(u_1)\mu_2(K) + o(h_1^2) + o(h_2^d)\\ &\quad + n^{-2}\sum_{i=1}^n\sum_{j=1}^n p(x^i)^{-1} e_1^T S^{-1}(x^i) A_j(x^i) W(X_{2i})\\ &\qquad \times\begin{pmatrix}1\\(U_{j1}-u_1)/h_1\\X_{3j}\end{pmatrix}\varepsilon_j + o_p(n^{-1/2}).\end{aligned}$$

By the projection argument which we used when treating $T_{n,1}$ in the proof of Theorem 1, we obtain

(6.25)
$$\hat{g}_1^+(u_1) - g_1^+(u_1) = \tfrac{1}{2}h_1^2 g_1''(u_1)\mu_2(K) + n^{-1}\sum_{j=1}^{n} K_{h_1}(U_{1j} - u_1)\varepsilon_j^* + o_p\big((nh_1)^{-1/2}\big),$$

where

$$\varepsilon_j^* = \frac{\varepsilon_j p_2(X_{2j}) e_1^T S^{-1}(x^j) W(X_{2j})}{p(x^j)} \begin{pmatrix} 1 \\ (U_{j1} - u_1)/h_1 \\ X_{3j} \end{pmatrix}.$$

Therefore, by checking the Lyapounov condition, we can establish Theorem 5, where the variance is obtained from (6.25) along with some algebra.

Acknowledgments. The manuscript was completed while Fan was visiting the Department of Statistics, the Chinese University of Hong Kong, and he is grateful for their hospitality. Furthermore, the authors would like to thank S. Sperlich for computational assistance and S. Profit for helpful remarks.

REFERENCES

BERNDT, E. R. (1991). *The Practice of Econometrics: Classic and Contemporary*. Addison-Wesley, Reading, MA.

BHATTACHARYA, P. K. and ZHAO, P.-L. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.* **25** 244–262.

BUJA, A., HASTIE, T. J. and TIBSHIRANI, R. J. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–510.

CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489.

CHEN, R., HÄRDLE, W., LINTON, O. and SEVERANCE-LOSSIN, E. (1996). Estimation and variable selection in additive nonparametric regression models. In *Statistical Theory and Computational Aspects of Smoothing* (W. Härdle and M. Schimek, eds.). Physika, Heidelberg.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21** 196–216.

FAN, J. (1997). Comments on "Polynomial splines and their tensor product in the extended linear models" by C. J. Stone, M. H. Hansen, C. Kooperberg and Y. U. Troung. *Ann. Statist.* **25** 1425–1432.

FAN, J. and GIBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008–2036.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

FRANZ, W. (1991). *Arbeitsökonomik*. Springer, Berlin.

GASSER, T. and MÜLLER, H.-G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation. Lecture Notes in Math.* **757** 23–68. Springer, New York.

HÄRDLE, W. and MAMMEN, E. (1993). Testing parametric versus nonparametric regression. *Ann. Statist.* **21** 1926–1947.

HÄRDLE, W., MAMMEN, E. and MÜLLER, M. (1995). Testing parametric versus semiparametric modelling in generalized linear models. Technical Report.

HÄRDLE, W. and TSYBAKOV, A. B. (1995). Additive nonparametric regression on principal components, *J. Nonparametr. Statist.* **5** 157–184.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

HENGARTNER, N. W. (1996). Rate optimal estimation of additive regression via the integration method in the presence of many covariates. Unpublished manuscript.

LINTON, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika* **84** 469–473.

LINTON, O. B., MAMMEN, E. and NIELSEN, J. P. (1997). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Preprint.

LINTON, O. B. and NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82** 93–101.

MACK, Y. P. and SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61** 405–415.

OPSOMER, J. D. (1997). On the existence and asymptotic properties of backfitting estimators. Preprint.

OPSOMER, J. D. and RUPPERT, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.* **25** 186–211.

RUPPERT, D. and WAND, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22** 1346–1370.

SPECKMAN, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–436.

STONE, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. In *Recent Advances in Statistics: Papers Presented in Honor of Herman Chernoff's Sixtieth Birthday* (M. H. Rizvi, J. S. Rustagi and D. Siegmund, eds.). Academic Press, New York.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 685–705.

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 592–606.

TJØSTHEIM, D. and AUESTAD, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89** 1398–1409.

TREIMAN, D. J. (1978). Probleme der Begriffsbildung und Operationalisierung in der international vergleichenden Mobilitätsforschung. In *Sozialstrukturanalysen mit Umfragedaten* (F. U. Pappi, ed.). Athenäum, Kronberg im Taunus.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599-3260

INSTITUT FÜR STATISTIK UND ÖKONOMETRIE
WIRTSCHAFTSWISSENSCHAFTLICHE FAKULTÄT
HUMBOLDT-UNIVERSITÄT ZU BERLIN
SPANDAUER STRASSE 1
10178 BERLIN
GERMANY

INSTITUT FÜR ANGEWANDTE MATHEMATIK
RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY
E-MAIL: mammen@statlab.uni-heidelberg.de